*DOI: 10.25140/2411-5363-2022-4(30)-130-139*
*УДК 629.7:519.2*

***Артем Волокита[1], Богдан Герега[2]***

[1]кандидат технічних наук, доцент, кафедра обчислювальної техніки
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (Київ, Україна)
**E-mail:** artem.volokita@kpi.ua. **ORCID:** http://orcid.org/0000-0001-9069-5544

[2]студент 6-го курсу факультету ІОТ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (Київ, Україна)
**E-mail:** bogdangerega19@gmail.com

# ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ СИСТЕМ ПРОГНОЗУВАННЯ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ

*Серцево-судинні захворювання щорічно вбивають близько 20,5 мільйонів людей. Раннє виявлення захворювання може допомогти людям змінити свій спосіб життя та забезпечити належне медичне лікування. У роботі представлені різні атрибути, пов'язані з хворобами серця, та модель на основі таких алгоритмів навчання: Logistic Regression, K-nearest neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier та XGBoost Classifier. Модель використовує набір даних із Клівлендської бази даних UCI для пацієнтів із серцевими захворюваннями. Набір даних містить 303 екземпляри та 76 атрибутів. З цих 76 атрибутів лише 14 атрибутів розглядаються для тестування, що важливо для обґрунтування продуктивності різних алгоритмів. Результати показують, що найвищий бал точності досягається з алгоритмом Support Vector Machine. Ця дослідницька робота має на меті продемонструвати ймовірність розвитку серцевих захворювань у пацієнтів використовуючи різні алгоритми машинного навчання.*

***Ключові слова:*** *машинне навчання; серцево-судинні захворювання; модель прогнозування; алгоритм; класифікація; регресія; набір даних; атрибут.*

*Рис.: 8. Табл.: 1. Бібл.: 14.*

**Актуальність теми дослідження.** Протягом останнього десятиліття серцево-судинні захворювання залишаються основною причиною смерті в усьому світі. За оцінками Всесвітньої організації охорони здоров'я, щороку в усьому світі від серцево- судинних захворювань помирає понад 17,9 мільйона, і з них 80 % – через ішемічну хворобу серця та церебральний інсульт [1]. Ефективна, точна та рання медична діагностика захворювань серця відіграє вирішальну роль у проведенні профілактичних заходів для запобігання смерті.

**Постановка проблеми.** Методи, які в цей час використовуються для прогнозування та діагностики захворювань серця, в першу чергу засновані на аналізі історії хвороби пацієнта, симптомів та звітів про фізичний огляд, які лікарі проводять. У більшості випадків медичним експертам важко точно передбачити захворювання серця у пацієнта, оскільки вони можуть передбачити з точністю до 67 %, оскільки наразі діагностика будь-якого захворювання проводиться за схожими симптомами від раніше діагностованих пацієнтів [2]. Отже, медична сфера потребує автоматизованої інтелектуальної системи для точного прогнозування захворювань серця. Цього можна досягти, використовуючи величезну кількість даних про пацієнтів, які доступні в медичній галузі, а також алгоритми машинного навчання.

**Аналіз останніх досліджень і публікацій.** У ході дослідження було здійснено аналіз чотирьох статтей, у яких розглядаються найсучасніші методи діагностики захворювань серця з використанням методів машинного навчання, які були здійснені в результаті різноманітних дослідницьких робіт. Розглянемо кожну з них:

1. R. Perumal та Kaladevi AC [3] розробили модель прогнозування серцево-судинних захворювань, використовуючи набір даних Клівленда з 303 екземплярів даних за допомогою стандартизації та зменшення ознак за допомогою PCA (principal component analysis), де вони визначили та використали сім основних компонентів для навчання класифікаторів ML. Вони дійшли висновку, що LR (Logistic regression) і SVM (Support-vector machine) забезпечили майже подібні значення точності (87 і 85 % відповідно) порівняно з k-NN (K-Nearest Neighbor) – 69 %.

2. Christalin Latha і Carolin Jeeva [4] провели порівняльний аналіз для підвищення точності прогнозування ризику серцево-судинних захворювань з використанням ансамблевих методів (ensemble methods) на наборі даних Клівленда з 303 спостережень. Вони застосували метод грубої сили (brute-force) для отримання всіх можливих комбінацій наборів атрибутів і навчили класифікатори. Вони досягли максимального збільшення точності класифікатора на основі алгоритму ансамблю та отримали точність 85,48%, використовуючи такі класифікатори NB (Naive Bayes), BN (Bayesian Network), RF (Random forest) та MLP (Multilayer perceptron) та використовуючи набір із дев'яти атрибутів.

3. Ananey-Obiri, Daniel і Sarku, Enoch [5] розробили три моделі класифікації, а саме, LR (Logistic regression), DT (Decision Tree) і GNB (Gaussian Naive Bayes), для прогнозування серцевих захворювань на основі набору даних Клівленда. Зменшення ознак було виконано за допомогою декомпозиції одного значення, яка зменшила ознаки з 13 до 4. Вони дійшли висновку, що як LR, так і GNB мали кращі прогнозні показники 82,75 %, а для DT трохи менше за 79,31 %.

4. Kumar, Sindhu та ін. [6] навчили п'ять класифікаторів машинного навчання, а саме, LR, SVM, DT, RF і KNN, використовуючи набір даних UCI з 303 записами та 10 атрибутами для прогнозування серцево-судинних захворювань. Класифікатор RF досяг найвищої точності 85,71 % порівняно з іншими класифікаторами (DT – 74,28 %, LR – 74,28 %, SVM – 77,14 %, K-NN – 68,57 %).

**Виділення недосліджених частин загальної проблеми.** Провівши аналіз останніх досліджень було визначено найефективніші алгоритми машинного навчання - Logistic Regression, K-nearest neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier та XGBoost Classifier. Також аналіз показав, що варіювання кількості атрибутів при прогнозуванні залишається недослідженою областю.

**Метою статті** є дослідження різних методів аналізу даних, корисних для ефективного прогнозування серцевих захворювань та розробка ефективного та точного підходу прогнозування з оптимальною кількістю атрибутів.

**Виклад основного матеріалу.** Важливим початковим етапом будь-якого прогнозування є створення або знаходження набору даних та подальша його обробка. Набір даних який використовувався у дослідженні - це дані зібрані у Клівленді про хворобу серця у форматі .csv із репозиторію машинного навчання UCI [7]. Потім вони були імпортовані у програмний інструмент, досліджувались атрибути, типи, діапазони значень та інша статистична інформація. Наступним кроком була попередня обробка даних, яка включала такі завдання, як пошук відсутніх значень у наборі даних та заміна відсутніх значень або константою користувача, або середнім значенням залежно від типу атрибута, щоб переконатися, що класифікатори машинного навчання забезпечують кращу продуктивність.

Набір даних Cleveland складається з 303 екземплярів із 76 атрибутами, але лише 14 атрибутів розглядаються для експериментальних цілей дослідження. Описи атрибутів для набору даних Cleveland наведено в таблиці.

Таблиця

*Опис атрибутів для набору даних Cleveland із репозиторію МН UCI*

| Атрибут | Опис | Тип атрибута | Діапазон значень атрибута |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| age | Вік у роках | Numeric | 29 - 77 |
| sex | Стать | Nominal | 0 - жінка,<br>1 - чоловік |
| cp | Тип болю в грудях | Nominal | 0 - типова стенокардія,<br>1 - нетипова стенокардія,<br>2 - біль без стенокардії,<br>3 - без симптомів |
| trestbps | Артеріальний тиск спокою в мм рт.ст. на момент оформлення у стаціонар | Numeric | 94 - 200 |

Закінчення табл.

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| chol | Сироватковий холестерин в mg/dL | Numeric | 125 - 564 |
| fbs | Цукор у крові (натще) > 120 mg/dL | Nominal | 0 - false, 1 - true |
| restecg | Результати електрокардіографії у стані спокою | Nominal | 0 - норма,<br>1 - аномалія ST-T wave,<br>2 - визначена гіпертрофія лівого шлуночка за критеріями Естеса |
| thalach | Досягнута максимальна частота серцевих скорочень | Numeric | 71 - 202 |
| exang | Вправи викликають стенокардію | Nominal | 0 - ні,<br>1 - так |
| oldpeak | Депресія ST, спричинена фізичними навантаженнями, порівняно зі спокоєм | Numeric | 0 - 6.2 |
| slope | Нахил піку вправи сегмент ST | Nominal | 0 - висхідний,<br>1 - плоский,<br>2 - низхідний |
| ca | Кількість магістральних судин, забарвлених за допомогою флюороскопії | Nominal | 0–4 |
| thal | Стан серця | Nominal | 0 -3 |
| target | Передбачуваний атрибут | Nominal | 0 - відсутність ризику серцево-судинних захворювань,<br>1 - ризик серцево-судинних захворювань |

Перед навчанням моделі було розглянуто та проаналізовано дані. Мета тут полягає в тому, щоб дізнатися більше про дані, з якими доводиться працювати. Було проведено аналіз даних, який включає статистику, яка підсумовує центральну тенденцію, дисперсію та форму розподілу набору даних, за винятком NaN значень. Результати наведено нижче (рис. 1).

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 | 303.00 |
| mean | 54.37 | 0.68 | 0.97 | 131.62 | 246.26 | 0.15 | 0.53 | 149.65 | 0.33 | 1.04 | 1.40 | 0.73 | 2.31 | 0.54 |
| std | 9.08 | 0.47 | 1.03 | 17.54 | 51.83 | 0.36 | 0.53 | 22.91 | 0.47 | 1.16 | 0.62 | 1.02 | 0.61 | 0.50 |
| min | 29.00 | 0.00 | 0.00 | 94.00 | 126.00 | 0.00 | 0.00 | 71.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 25% | 47.50 | 0.00 | 0.00 | 120.00 | 211.00 | 0.00 | 0.00 | 133.50 | 0.00 | 0.00 | 1.00 | 0.00 | 2.00 | 0.00 |
| 50% | 55.00 | 1.00 | 1.00 | 130.00 | 240.00 | 0.00 | 1.00 | 153.00 | 0.00 | 0.80 | 1.00 | 0.00 | 2.00 | 1.00 |
| 75% | 61.00 | 1.00 | 2.00 | 140.00 | 274.50 | 0.00 | 1.00 | 166.00 | 1.00 | 1.60 | 2.00 | 1.00 | 3.00 | 1.00 |
| max | 77.00 | 1.00 | 3.00 | 200.00 | 564.00 | 1.00 | 2.00 | 202.00 | 1.00 | 6.20 | 2.00 | 4.00 | 3.00 | 1.00 |

*Рис. 1. Результати аналізу даних*

Дослідивши набір даних, можна помітити, що потрібно перетворити деякі категорійні змінні (а саме: age, trestbps, chol, thalach, oldpeak) в фіктивні змінні та масштабувати всі значення перед навчанням моделей машинного навчання.

Дані було поділено на навчальний і тестовий набори, вони відповідно становлять 70 та 30 %.

У роботі застосовано кілька алгоритмів машинного навчання до одного набору даних, щоб визначити найкращий класифікатор для прогнозування захворювань. У цій роботі розглядаються різні класифікатори, які були використані для прогнозування ризику серцевих захворювань. Розглянемо послідовно кожен із них.

Логістична регресія (logistic regression) – статистичний регресійний метод, що застосовують у випадку, коли залежна змінна є бінарною, тобто може набувати тільки двох значень (0 або 1). Логістична регресія, по суті, використовує логістичну функцію для моделювання двійкової вихідної змінної, її вигляд такий:

$$\frac{1}{1+e^{-x}}. \quad (1)$$

LR не вимагає лінійного відношення між вхідними та вихідними змінними. LR – це рівняння, де кожен предиктор множиться на коефіцієнт і підсумовується. Ця сума стає аргументом для логістичної функції для передбачення класу. Для одного спостереження $x$ з $n$ ознаками відповідь $y$ визначається як:

$$\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}}. \quad (2)$$

Математика логістичної регресії спирається на концепцію «шансів» події, яка являє собою ймовірність того, що подія відбудеться, поділену на ймовірність того, що подія не відбудеться. Так само, як і в лінійній регресії, логістична регресія має ваги, пов'язані з розмірами вхідних даних, де зв'язок між ваговими коефіцієнтами та результатом моделі є експоненційним. Ця модель класифікації, яку дуже легко реалізувати, досягає дуже хорошої продуктивності за допомогою лінійно розділених класів [8].

Алгоритм K-найближчих сусідів (K-NN) є методом контрольованого алгоритму класифікації (supervised classification algorithm). Він класифікує об'єкти залежно від найближчого сусіда. Це тип навчання на основі прикладів (instance-based learning). Розрахунок відстані атрибута від його сусідів вимірюється за допомогою евклідової відстані. Алгоритм використовує групу іменованих точок та використовує їх, щоб позначити іншу точку. Дані групуються на основі подібності між ними, і можна заповнити відсутні значення даних за допомогою K-NN. Після заповнення відсутніх значень до набору даних застосовуються різні методи прогнозування. Можна отримати кращу точність, використовуючи різні комбінації. K-NN простий у виконанні. Цей алгоритм є універсальним і використовується для класифікації, регресії та пошуку [9].

Метод опорних векторів (support-vector machine) – це лінійна модель для задач класифікації та регресії. Він може вирішувати лінійні та нелінійні задачі та добре працювати для багатьох практичних задач. Ідея SVM проста: алгоритм створює лінію або гіперплощину, яка розділяє дані на класи. Відповідно до алгоритму SVM ми знаходимо точки, найближчі до прямої з обох класів. Ці точки називають опорними векторами. Тепер обчислюємо відстань між прямою та опорними векторами. Ця відстань називається маржею. Основна мета – максимізувати маржу. Гіперплощина, для якої запас є максимальним, є оптимальною гіперплощиною. Таким чином, SVM намагається прийняти межу рішення таким чином, щоб поділ між двома класами був якомога ширшим [10].

Дерево рішень (Decision Tree) - це алгоритм класифікації, який працює як з категоріальними, так і з числовими даними. Дерево рішень використовується для створення деревоподібних структур. Дерево рішень просте і широко використовується для обробки медичних даних. Легко реалізувати та проаналізувати дані у вигляді деревоподібного графіка. Модель дерева рішень проводить аналіз на основі трьох вузлів. Кореневий вузол: головний вузол, на основі якого функціонують усі інші вузли. Внутрішній вузол: обробляє різні атрибути. Листковий вузол: представляє результат кожного тесту. Цей алгоритм розбиває дані на два або більше аналогічних набори на основі найважливіших показників. Обчислюється ентропія кожного атрибута, а потім дані поділяються, причому предикатори мають максимальний приріст інформації або мінімальну ентропію. Алгоритм може бути дуже корисним для вирішення проблем, пов'язаних із прийняттям рішень [11].

Алгоритм випадкового лісу (Random forest) — це алгоритм керованого навчання (supervised learning algorithm). У цьому алгоритмі кілька дерев створюють ліс. Кожне окреме дерево у випадковому лісі дає очікування класу, а клас з найбільшою кількістю голосів перетворюється на прогноз моделі. У класифікаторі Random forest більша кількість дерев дає вищу точність, наприклад. Алгоритм використовується як для задач класифікації, так і для завдань регресії. Він працює в чотири кроки:

1) Вибираються випадкові вибірки з заданого набору даних.

2) Будується дерево рішень для кожної вибірки та отримуємо результат прогнозу з кожного дерева рішень.

3) Проводиться голосування за кожен прогнозований результат.

4) Вибирається результат прогнозу з найбільшою кількістю голосів як остаточний прогноз.

Random forest вважається дуже точним і надійним методом через кількість дерев рішень, які беруть участь у процесі. Проте алгоритм повільно генерує прогнози, оскільки має кілька дерев рішень [12].

Алгоритм екстремального градієнтного підсилювання (eXtreme Gradient Boosting) - це реалізація Random forest із покращеним градієнтом, розроблена для більшої швидкості та продуктивності, що є домінуючим фактором машинного навчання. XGBoost належить до сімейства алгоритмів підсилювання (boosting algorithms) і використовує в своїй основі структуру підвищення градієнта (GBM). Основна ідея, що лежить в основі алгоритмів підсилювання, полягає в тому, щоб створити відносно слабку модель, зробити висновки про важливість і параметри різних функцій, а потім використати ці висновки для побудови нової, сильнішої моделі та отримати вигоду з помилкової класифікації попередньої моделі та спробувати її зменшити. Слід знати про базові програми XGBoost, які навчаються за замовчуванням: деревні ансамблі (tree ensembles). Модель деревного ансамблю – це набір дерев класифікації та регресії (CART). Дерева вирощують одне за одним, і в наступних ітераціях робляться спроби зменшити рівень помилкової класифікації. Кожне дерево дає різну оцінку прогнозу залежно від даних, які воно бачить, і бали кожного окремого дерева підсумовуються, щоб отримати остаточний бал. XGBoost реалізує паралельну обробку і це дозволяє змешити час отримання рішення. XGBoost є потужним алгоритм машинного навчання, особливо коли йдеться про швидкість та точність [13].

Для аналізу алгоритмів машинного навчання було використано sklearn, що є безкоштовною програмною бібліотекою машинного навчання для мови програмування Python, яка надає функціональність для створення та тренування різноманітних алгоритмів класифікації, регресії та кластеризації. Бібліотека містить реалізовані класифікатори такі як: LogisticRegression, KNeighborsClassifier, SVC, DecisionTreeClassifier та RandomForestClassifier. А також було використано бібліотеку xgboost для використання класифікатора XGBClassifier.

Основними метриками, на основі яких визначалась ефективність алгоритмів є: accuracy score (оцінка точності класифікації), confusion matrix (матриця невідповідностей) та classification report (текстовий звіт з основними показниками класифікації).

Результати метрик для моделі LR наведені на рисунку 2. Можна помітити, що точності в навчальному та тестовому наборі практично однакові. Матриця невідповідностей показує, що в навчальному наборі до класу 0 (відсутність ризику ССЗ) віднесено 80 екземплярів, а згідно з Classification Report має бути 97 екземплярів, та до класу 1 (ризик ССЗ) - 104 зі 115 екземплярів. У тестовому наборі ситуація наступна: до класу 0 – 34 з 41 екземплярів, до класу 1 – 45 з 50.

Train Result:

Accuracy Score: 86.79%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.88 | 0.86 | 0.87 | 0.87 | 0.87 |
| recall | 0.82 | 0.90 | 0.87 | 0.86 | 0.87 |
| f1-score | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| support | 97.00 | 115.00 | 0.87 | 212.00 | 212.00 |

Confusion Matrix:

```
[[ 80  17]
 [ 11 104]]
```

Test Result:

Accuracy Score: 86.81%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.87 | 0.87 | 0.87 | 0.87 | 0.87 |
| recall | 0.83 | 0.90 | 0.87 | 0.86 | 0.87 |
| f1-score | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| support | 41.00 | 50.00 | 0.87 | 91.00 | 91.00 |

Confusion Matrix:

```
[[34  7]
 [ 5 45]]
```

*Рис. 2. Результати метрик для моделі LR*

Результати для моделі K-NN наведені на рисунку 3. Наочно видно, що значення точності на навчальному наборі та тестовому наборі майже ідентичні. Матриця невідповідностей показує, що 82 екземпляри в навчальному наборі були віднесені до класу 0, тоді як згідно зі звітом про класифікацію має бути 97 випадків, а 102 із 115 випадків були віднесені до класу 1. У тестовому наборі ми отримали, що до класу 0 віднесено 35 з 41 екземплярів, до класу 1 – 44 з 50 екземплярів.

Train Result:

Accuracy Score: 86.79%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.86 | 0.87 | 0.87 | 0.87 | 0.87 |
| recall | 0.85 | 0.89 | 0.87 | 0.87 | 0.87 |
| f1-score | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| support | 97.00 | 115.00 | 0.87 | 212.00 | 212.00 |

Confusion Matrix:

```
[[ 82  15]
 [ 13 102]]
```

Test Result:

Accuracy Score: 86.81%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| recall | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| f1-score | 0.85 | 0.88 | 0.87 | 0.87 | 0.87 |
| support | 41.00 | 50.00 | 0.87 | 91.00 | 91.00 |

Confusion Matrix:

```
[[35  6]
 [ 6 44]]
```

*Рис. 3. Результати метрик для моделі K-NN*

Результати для моделі SVC наведені на рисунку 4. Точності в начальному та тестовому наборах є вищими, ніж у моделях LR та K-NN. Згідно з матрицею невідповідності у навчальному наборі до класу 0 віднесено 89 з 97 екземплярів, та до класу 1 – 109 зі 115 екземплярів. У тестовому наборі ситуація така: до класу 0 – 36 з 41 екземплярів, до класу 1 – 44 з 50.

Train Result:

Accuracy Score: 93.40%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.94 | 0.93 | 0.93 | 0.93 | 0.93 |
| recall | 0.92 | 0.95 | 0.93 | 0.93 | 0.93 |
| f1-score | 0.93 | 0.94 | 0.93 | 0.93 | 0.93 |
| support | 97.00 | 115.00 | 0.93 | 212.00 | 212.00 |

Confusion Matrix:

```
[[ 89   8]
 [  6 109]]
```

Test Result:

Accuracy Score: 87.91%

CLASSIFICATION REPORT:

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.86 | 0.90 | 0.88 | 0.88 | 0.88 |
| recall | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 |
| f1-score | 0.87 | 0.89 | 0.88 | 0.88 | 0.88 |
| support | 41.00 | 50.00 | 0.88 | 91.00 | 91.00 |

Confusion Matrix:

```
[[36  5]
 [ 6 44]]
```

*Рис. 4. Результати метрик для моделі SVC*

Результати для моделі DT наведені на рисунку 5. Точність у навчальному наборі становить 100 %, що є найвищим показником. У тестовому наборі точність дорівнює 78,02 %, що є меншим значенням, ніж у попередніх моделях. Матриця невідповідностей показує, що всі екземпляри вірно класифіковано у навчальному наборі, а у тестовому ситуація така: до класу 0 віднесено 34 з 41 екземплярів, до класу 1 – 37 з 50.

```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  1.00   1.00      1.00       1.00          1.00
recall     1.00   1.00      1.00       1.00          1.00
f1-score   1.00   1.00      1.00       1.00          1.00
support   97.00 115.00      1.00     212.00        212.00
________________________________________________________
Confusion Matrix:
 [[ 97   0]
 [  0 115]]
```

```
Test Result:
================================================
Accuracy Score: 78.02%
________________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision  0.72  0.84      0.78       0.78          0.79
recall     0.83  0.74      0.78       0.78          0.78
f1-score   0.77  0.79      0.78       0.78          0.78
support   41.00 50.00      0.78      91.00         91.00
________________________________________________________
Confusion Matrix:
 [[34  7]
 [13 37]]
```

*Рис. 5. Результати метрик для моделі DT*

Результати для моделі RF наведені на рисунку 6. Точність у навчальному наборі складає максимум - 100%, а у тестовому наборі 82.42%, що є меншим значенням ніж у LR та K-NN, проте більшим, ніж у DT. Згідно матриці невідповідності всі екземпляри навчального набору вірно класифіковано. Натомість у тестовому наборі до класу 0 віднесено 33 з 41 екземплярів, до класу 1 – 42 з 50.

```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  1.00   1.00      1.00       1.00          1.00
recall     1.00   1.00      1.00       1.00          1.00
f1-score   1.00   1.00      1.00       1.00          1.00
support   97.00 115.00      1.00     212.00        212.00
________________________________________________________
Confusion Matrix:
 [[ 97   0]
 [  0 115]]
```

```
Test Result:
================================================
Accuracy Score: 82.42%
________________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision  0.80  0.84      0.82       0.82          0.82
recall     0.80  0.84      0.82       0.82          0.82
f1-score   0.80  0.84      0.82       0.82          0.82
support   41.00 50.00      0.82      91.00         91.00
________________________________________________________
Confusion Matrix:
 [[33  8]
 [ 8 42]]
```

*Рис. 6. Результати метрик для моделі RF*

Результати для моделі XGB наведені на рисунку 7. На навчальному наборі ми отримали точність 98.58%, а на тестовому - 83.52%, що є меншим значенням, ніж у LR та K-NN, проте більшим, ніж у DT та RF. При визначенні класу ризику ССЗ 95 з 97 екземплярів вірно класифіковано згідно матриці невідповідностей та до класу ризик ССЗ - 114 зі 115 екземплярів. У тестовому наборі ситуація наступна: до класу 0 – 34 з 41 екземплярів, до класу 1 – 42 з 50.

```
Train Result:
================================================
Accuracy Score: 98.58%
________________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  0.99   0.98      0.99       0.99          0.99
recall     0.98   0.99      0.99       0.99          0.99
f1-score   0.98   0.99      0.99       0.99          0.99
support   97.00 115.00      0.99     212.00        212.00
________________________________________________________
Confusion Matrix:
 [[ 95   2]
 [  1 114]]
```

```
Test Result:
================================================
Accuracy Score: 83.52%
________________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision  0.81  0.86      0.84       0.83          0.84
recall     0.83  0.84      0.84       0.83          0.84
f1-score   0.82  0.85      0.84       0.83          0.84
support   41.00 50.00      0.84      91.00         91.00
________________________________________________________
Confusion Matrix:
 [[34  7]
 [ 8 42]]
```

*Рис. 7. Результати метрик для моделі XGB*

Результати точності для всіх моделей наведені на рис. 8. З досліджень наглядно видно, що найкращу точність на тестовому наборі має модель Support Vector Machine – 87,91 %, а найнижчу Decision Tree Classifier – 78,02 %.

| | Model | Training Accuracy % | Testing Accuracy % |
|---|---|---|---|
| 0 | Logistic Regression | 86.79 | 86.81 |
| 1 | K-nearest neighbors | 86.79 | 86.81 |
| 2 | Support Vector Machine | 93.40 | 87.91 |
| 3 | Decision Tree Classifier | 100.00 | 78.02 |
| 4 | Random Forest Classifier | 100.00 | 82.42 |
| 5 | XGBoost Classifier | 98.58 | 83.52 |

*Рис. 8. Результати точності для всіх моделей*

У результаті дослідження було виявлено, що зменшення кількості атрибутів не завжди призводить до підвищення ефективності знаходження вірного рішення.

**Висновки.** Загальна мета дослідження полягала у визначенні різних методів аналізу даних, корисних для ефективного прогнозування серцевих захворювань та здійснення ефективного і точного прогнозування з оптимальною кількістю атрибутів. У цьому дослідженні було розглянуто 14 основних атрибутів, було застосовано шість методів класифікації: Logistic Regression, K-nearest neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier та XGBoost Classifier. Дані були попередньо оброблені, а потім використані в моделях.

Загалом, жоден алгоритм не «кращий» за інший. Існує теорема «No Free Lunch» [14]. У ній стверджується, що будь-які два алгоритми оптимізації є еквівалентними, якщо їх продуктивність усереднена для всіх можливих проблем. Проте з практичної точки зору можна стверджувати, що Logistic Regression показує кращу точність, тому що він дуже швидко класифікує невідомі записи та є менш схильним до перенавчання (overfitting). Що стосується K-nearest neighbors, то у випадку нелінійних даних класифітор є дуже ефективним і забезпечує високу точність, також він є стійким до зашумлених навчальних даних (noisy training data). Support Vector Machine є ефективним на невеликих наборах даних, SVM надає налаштування дуже корисного параметра kernel і за допомогою застосування відповідної функції ядра ми можемо вирішити будь-яку складну проблему. Також SVM зазвичай не страждає від пере- навчання та у порівнянні з іншими класифікаторами має кращу обчислювальну складність, і навіть якщо кількість позитивних та негативних прикладів неоднакова, можна використовувати SVM, оскільки він має можливість нормалізації. Дані переваги класифікатора SVM, згідно досліджень, забезпечили найкращу точність, яка становить 87,91 % на тестовому наборі.

## Список використаних джерел

1. World Health Organization // Cardiovascular diseases. – World Health Organization. 2022 [Electronic resource]. – Accessed mode: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

2. Ansari M. F. A prediction of heart disease using machine learning algorithms / Ansari M. F., Alankar B., Kaur H. // International conference on image processing and capsule networks. – 2020. – Pp. 497-504.

3. Ramya Perumal. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques [Electronic resource] / Ramya Perumal, Kaladevi AC. // International Journal of Advanced Science and Technology. – 2020. – Vol. 29(06). – Pp. 4225-4234. – Accessed mode: http://sersc.org/journals/index.php/IJAST/article/view/16428.

4. Latha C. B. C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques / Latha C. B. C., Jeeva S. C. // Informatics in Medicine Unlocked. – 2019. – Vol. 16. – Atr. 100203.

5. Ananey-Obiri D. Predicting the presence of heart diseases using comparative data mining and machine learning algorithms / Ananey-Obiri D., Sarku E. // International Journal of Computer Applications. – 2020. – Vol. 176(11). – Pp. 17-21.

6. Kumar N. K. Analysis and prediction of cardio vascular disease using machine learning classifiers / Kumar N. K., Sindhu G. S., Prashanthi D. K., Sulthana A. S. // 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE. – 2020, March. – Pp. 15-21.

7. UCI Machine Learning Repository [Electronic resource] / Janosi A., Steinbrunn W., Pfisterer M., Detrano R. – 1988.– Accessed mode: https://archive-beta.ics.uci.edu/ml/datasets/heart+disease.

8. Maini E. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India / Maini E., Venkateswarlu B., Maini B., Marwaha D. // Medical Journal Armed Forces India. – 2021. – Vol. 77(3). – Pp. 302-311.

9. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease / Pouriyeh S., Vahid S., Sannino G., De Pietro G., Arabnia H., Gutierrez J. // 2017 IEEE symposium on computers and communications (ISCC). – 2017, July. – Pp. 204-207.

10. Support vector machine classification of microarray data / Mukherjee S., Tamayo P. A. S. D., Slonim D., Verri A., Golub T., Mesirov J., Poggio T. // AI Memo. – 1677, Massachusetts Institute of Technology, 1999.

11. Priyam A. Comparative analysis of decision tree classification algorithms / Priyam A., Abhijeeta G. R., Rathee A., Srivastava S. // International Journal of current engineering and technology. – 2013. – Vol. 3(2). – Pp. 334-337.

12. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework / Xu S., Zhang Z., Wang D., Hu J., Duan X., Zhu T. // 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). – 2017, March. – Pp. 228-232.

13. Chen T. Xgboost: A scalable tree boosting system / Chen T., Guestrin C. // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016, August. – Pp. 785-794.

14. Wolpert D. H. No free lunch theorems for optimization / Wolpert D. H., Macready W. G. // IEEE transactions on evolutionary computation. – 1997. – Vol. 1(1). – Pp. 67-82.

## References

1. World Health Organization. (2022). Cardiovascular diseases. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

2. Ansari, M. F., Alankar, B., & Kaur, H. (2020, May). A prediction of heart disease using machine learning algorithms. In International conference on image processing and capsule networks (pp. 497-504). Springer, Cham.

3. Ramya Perumal, Kaladevi AC. (2020). Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. *International Journal of Advanced Science and Technology, 29*(06), 4225-4234. http://sersc.org/journals/index.php/IJAST/article/view/16428.

4. Latha, C.B.C., & Jeeva, S.C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked, 16*, 100203.

5. Ananey-Obiri, D., & Sarku, E. (2020). Predicting the presence of heart diseases using comparative data mining and machine learning algorithms. *International Journal of Computer Applications, 176*(11), 17-21.

6. Kumar, N.K., Sindhu, G.S., Prashanthi, D.K., & Sulthana, A.S. (2020, March). Analysis and prediction of cardio vascular disease using machine learning classifiers. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 15-21). IEEE.

7. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R. (1988). UCI Machine Learning Repository. https://archive-beta.ics.uci.edu/ml/datasets/heart+disease.

8. Maini, E., Venkateswarlu, B., Maini, B., & Marwaha, D. (2021). Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. *Medical Journal Armed Forces India, 77*(3), 302-311.

9. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE symposium on computers and communications (ISCC)* (pp. 204-207). IEEE.

10. Mukherjee, S., Tamayo, P. A. S. D., Slonim, D., Verri, A., Golub, T., Mesirov, J., & Poggio, T. (1999). Support vector machine classification of microarray data. *AI Memo 1677,* Massachusetts Institute of Technology.

11.Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology, 3*(2), 334-337.

12.Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X., & Zhu, T. (2017, March). Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (pp. 228-232). IEEE.

13.Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

14.Wolpert, DH., & Macready, W.G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation, 1*(1), 67-82.

Отримано 25.11.2022

*UDC 629.7:519.2*

***Artem Volokita[1], Bohdan Hereha[2]***

[1]PhD in Technical Sciences, Associate Professor, Department of Computer Engineering
National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute" (Kyiv, Ukraine)
**E-mail:** artem.volokita@kpi.ua. **ORCID:** http://orcid.org/0000-0001-9069-5544

[2]student of the 6th year of the Faculty of IOT
National Technical University of Ukraine "Kyiv Polytechnic Institute
named after Igor Sikorsky" (Kyiv, Ukraine)
**E-mail:** bogdangerega19@gmail.com

# COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR DISEASES PREDICTION SYSTEMS

*Every year, around 20.5 million people die from cardiovascular disorders. People who are diagnosed with the disease early can alter their lifestyles and receive appropriate medical care. A model based on the learning algorithms Logistic Regression, K-nearest Neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, and XGBoost is presented in the paper along with various heart disease-related attributes. The Cleveland UCI database of heart disease patients is used in the model. There are 303 instances and 76 attributes in the data set. Only 14 of these 76 attributes—which are crucial to justifying the effectiveness of various algorithms—are taken into account during testing. The main contribution of this research work is the implementation of an intuitively understandable system of medical forecasts for the diagnosis of heart diseases using modern methods of machine learning. Algorithms used for predicting heart diseases are discussed in this work, and a comparison is made between existing systems. Six classification methods were used: Logistic Regression, K-nearest neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier. 14 attributes were used to predict cardiovascular disease, which is a much better solution than using 5 or 10, as was the case in the reviewed papers. To ensure high accuracy, hyperparameters were adjusted for each classifier. As a result, good performance was obtained. In this work, the SVM classifier proved to be the most effective, providing an accuracy of 87.91 % on the test set. It was possible to achieve greater accuracy than in the studied works.*

***Keywords:*** *machine learning; cardiovascular disease; prediction model; algorithm; classification; regression; dataset; attribute.*

*Fig.: 8. Table: 1. References: 14.*

Волокита А., Герега Б. Порівняльний аналіз алгоритмів машинного навчання для систем прогнозування серцево-судинних захворювань. *Технічні науки та технології.* 2022. № 4(30). С. 130-139.